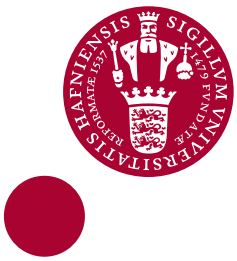

Differential diagnosis of mild cognitive impairment and Alzheimer's disease using structural MRI cortical thickness, hippocampal shape, hippocampal texture, and volumetry

Sørensen, Lauge; Igel, Christian; Pai, Akshay Sadananda Uppinakudru; Balas, Ioana; Anker, Cecilie; Lillholm, Martin; Nielsen, Mads

Published in:
NeuroImage: Clinical

DOI:
10.1016/j.nicl.2016.11.025

Publication date:
2017

Document version
Publisher's PDF, also known as Version of record

NeuroImage: Clinical 13 (2017) 470–482

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

Differential diagnosis of mild cognitive impairment and Alzheimer's disease using structural MRI cortical thickness, hippocampal shape, hippocampal texture, and volumetry☆

Lauge Sørensen[a,b,*], Christian Igel[a], Akshay Pai[a,b], Ioana Balas[a], Cecilie Anker[b], Martin Lillholm[a,b], Mads Nielsen[a,b], for the Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing

[a]Department of Computer Science, University of Copenhagen, Copenhagen Ø DK-2100, Denmark
[b]Biomediq A/S, Copenhagen Ø DK-2100, Denmark

ARTICLE INFO

Article history:
Received 10 July 2016
Received in revised form 21 October 2016
Accepted 26 November 2016
Available online 7 December 2016

Keywords:
Alzheimer's disease
Biomarker
Classification
Machine learning
Mild cognitive impairment
Structural MRI

ABSTRACT

This paper presents a brain T1-weighted structural magnetic resonance imaging (MRI) biomarker that combines several individual MRI biomarkers (cortical thickness measurements, volumetric measurements, hippocampal shape, and hippocampal texture). The method was developed, trained, and evaluated using two publicly available reference datasets: a standardized dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the imaging arm of the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL). In addition, the method was evaluated by participation in the Computer-Aided Diagnosis of Dementia (CADDementia) challenge. Cross-validation using ADNI and AIBL data resulted in a multi-class classification accuracy of 62.7% for the discrimination of healthy normal controls (NC), subjects with mild cognitive impairment (MCI), and patients with Alzheimer's disease (AD). This performance generalized to the CADDementia challenge where the method, trained using the ADNI and AIBL data, achieved a classification accuracy 63.0%. The obtained classification accuracy resulted in a first place in the challenge, and the method was significantly better (McNemar's test) than the bottom 24 methods out of the total of 29 methods contributed by 15 different teams in the challenge. The method was further investigated with learning curve and feature selection experiments using ADNI and AIBL data. The learning curve experiments suggested that neither more training data nor a more complex classifier would have improved the obtained results. The feature selection experiment showed that both common and uncommon individual MRI biomarkers contributed to the performance; hippocampal volume, ventricular volume, hippocampal texture, and parietal lobe thickness were the most important. This study highlights the need for both subtle, localized measurements and global measurements in order to discriminate NC, MCI, and AD simultaneously based on a single structural MRI scan. It is likely that additional non-structural MRI features are needed to further improve the obtained performance, especially to improve the discrimination between NC and MCI.

1. Introduction

Structural magnetic resonance imaging (MRI) biomarkers of Alzheimer's disease (AD) are an active research area, and a wide range of biomarkers have been proposed and investigated (Ramani et al., 2006; Cuingnet et al., 2011; Falahati et al., 2014). The importance of structural MRI in AD was underlined by its inclusion in criteria for AD diagnosis (Jack et al., 2011a). To date, the volume of the hippocampus is the most studied and used structural MRI biomarker of AD (Jack et al., 2011b), and it is so far the only structural MRI biomarker that has been qualified for enrichment of clinical trials (Hill et al., 2014). Volumetry is generally a popular type of

☆ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) and from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at www.aibl.csiro.au.

* Corresponding author at: University of Copenhagen, Department of Computer Science, Universitetsparken 5, Copenhagen Ø DK-2100, Denmark.
E-mail address: lauges@diku.dk (L. Sørensen).

http://dx.doi.org/10.1016/j.nicl.2016.11.025

biomarker, and the region of interest (ROI) is not limited to the hippocampus. Examples of other ROIs include the amygdala (Poulin et al., 2011), the ventricles (Tanabe et al., 1997), and the whole brain (Tanabe et al., 1997). Other types of biomarkers include cortical thickness measurements (Singh et al., 2006; Eskildsen et al., 2013), shape (Gerardin et al., 2009; Achterberg et al., 2014), texture (Chincarini et al., 2011; Sørensen et al., 2016), proximity of brain structures (Lillemark et al., 2014), whole brain dissimilarities computed from a deformation (Klein et al., 2010), and methods based on voxel-wise modulated intensities (Ashburner and Friston, 2000; Davatzikos et al., 2008; Klöppel et al., 2008).

Some MRI biomarkers complement each other. For example, it has been shown that hippocampal shape and texture provide diagnostic information independent of hippocampal volume (Achterberg et al., 2014; Sørensen et al., 2016). Moreover, markers applied in different parts of the brain are expected to be sensitive to different stages of the disease. For example, the hippocampus is affected early by neurofibrillary tangles, a pathological hallmark of AD, whereas the cortex is only affected later (Braak and Braak, 1991). This has been empirically reflected by hippocampal volume being better at separating normal controls (NC) and mild cognitive impairment (MCI), the prodromal stage of AD (Colliot et al., 2008), whereas cortical thickness measurements have been shown to better separate MCI from AD (Singh et al., 2006). A combination of such complementary biomarkers may provide an overall better biomarker, especially when several diagnostic groups are considered (e.g., NC, MCI, AD). Accordingly, the combination of volumetry and cortical thinning has been used in several studies (Falahati et al., 2014).

Considering the multitude of MRI biomarkers, there is a need for standardized comparisons of methods on the same dataset in order to better understand how different biomarkers perform and what their relations are. The recent release of standardized datasets for the comparison of algorithms from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Wyman et al., 2013) was an important step in this direction, especially because ADNI is the most commonly used reference database in AD MRI biomarker research (Weiner et al., 2012). Likewise, recent large empirical comparison studies (Cuingnet et al., 2011; Sabuncu et al., 2015) and the Computer-Aided Diagnosis of Dementia (CADDementia) challenge (Bron et al., 2015) have been important steps.

CADDementia, hosted by the 17th International Conference on Medical Image Computing & Computer-Assisted Intervention (MICCAI), was a challenge on differential diagnosis of NC, subjects with MCI and patients suffering from AD based on a single structural MRI scan. The challenge provided an opportunity for different research groups to directly compare their methods in a completely standardized fashion. Notable characteristics of CADDementia included:

- differential diagnosis of NC, MCI, and AD, i.e., considering the three-class classification problem rather than pairwise comparisons, which are often reported in the literature;
- validation on a completely independent and unseen dataset from a different cohort; and
- standardized validation, i.e., evaluation metrics defined, implemented, and applied by the CADDementia team.

In this study, we propose a method that combines a range of volumetric measurements, cortical thickness measurements, hippocampal texture, and hippocampal shape, to obtain a combination biomarker that used more of the information contained in a structural MRI scan compared with a single biomarker approach. The purpose is simultaneous differential diagnosis of NC, MCI, and AD. This is a unique combination of basic MRI biomarkers not used before. The combination is achieved by entering all biomarkers as features in a linear discriminant analysis (LDA) classifier. The proposed method was developed and trained on a combination of MRI scans from ADNI (Wyman et al., 2013) and the MRI imaging arm of the Australian Imaging, Biomarker & Lifestyle Flagship Study of Aging (AIBL) (Ellis et al., 2009). The method was evaluated using cross-validation on the combination of ADNI and AIBL data, and by participation in the CADDementia challenge. The classification accuracy (CA) and area under the receiver operating characteristic (ROC) curve (AUC) estimated using the combination of ADNI and AIBL were comparable to the performance achieved in the CADDementia challenge, and the CADDementia test set CA was the highest among the participating teams, resulting in the first place at the challenge. Our method also achieved the highest AUC in the challenge.

A preliminary version of the work presented here appeared in the CADDementia workshop proceedings (Sørensen et al., 2014). The present study contains a detailed description of the method and the obtained results. Moreover, we perform additional experiments to investigate feature relations and importance in the method, and to investigate whether more data and/or a more complex classifier would have benefited the method.

2. Material and methods

2.1. The CADDementia challenge

The CADDementia challenge (Bron et al., 2015) used data collected from three different sites, and the data were split into two datasets, a validation dataset (30 scans) that included the clinical diagnosis, and a test data set (354 scans) for which clinical diagnosis was not available to the challenge participants. Participants were encouraged to use other data sources such as ADNI for training of their methods. The CADDementia validation set was mainly supplied for participants to gauge their performance. The CADDementia test set was the data to be analyzed and scores be submitted for the challenge.

The data was made available via the CADDementia website[1] on March 1, 2014, and test set scores were to be uploaded via the website no later than June 16, 2014. A total of 15 teams uploaded scores. Each team was allowed to upload 5 attempts, and a total of 29 attempts were uploaded. The results of the challenge were revealed at the challenge workshop at MICCAI on September 18, 2014.

The Python scripts used for evaluation of the results were also made available via the website, so each team could compute their performance on the validation set using the exact same code that would later be used to compute the performance on the test set.

Methods were ranked using the three-class CA (for NC, MCI, and AD), which was based on a hard classification output of the methods. Teams could supply soft classification outputs as well in order to also have ROC statistics computed. The AUC acted as secondary performance measure, i.e., ties according to CA were to be resolved using AUC.

2.2. Data

Data used in this study were obtained from three different cohorts: ADNI, AIBL, and CADDementia. Table 1 provides an overview.

2.2.1. ADNI standardized dataset

We used the "complete annual year 2 visits" 1.5T dataset from the collection of standardized datasets released by ADNI (Wyman et al., 2013). Raw unprocessed 1.5 T T1-weighted MRI images were downloaded from the ADNI database between February 1, 2012 and November 11, 2012, and the standardized dataset definition was downloaded from the ADNI website (http://adni.loni.usc.edu/methods/mri-analysis/adni-standardized-data/) on September 28, 2012.

[1] http://caddementia.grand-challenge.org/.

Table 1
Characteristics of the datasets.

	n	Age, years (mean ± std)	Male (%)	MMSE[a] (mean ± std)	Field strength (1.5 T/3 T)
ADNI standardized dataset					
All	504	75.3 ± 6.5	58.1	27.0 ± 2.6	504/0
NC	169	76.0 ± 5.1	50.9	29.2 ± 1.0	169/0
MCI	234	74.9 ± 7.0	66.7	27.1 ± 1.7	234/0
AD	101	75.3 ± 7.4	50.5	23.2 ± 1.9	101/0
ADNI HHP subset					
All	40	74.0 ± 7.6	47.5	26.3 ± 2.9	40/0
NC	13	75.9 ± 6.8	46.2	28.8 ± 1.1	13/0
MCI	11	70.4 ± 7.4	54.5	27.5 ± 1.2	11/0
AD	16	74.9 ± 8.0	43.8	23.4 ± 2.0	16/0
AIBL imaging arm					
All	145	75.4 ± 7.4	46.2	27.1 ± 4.0	1/144
NC	88	75.2 ± 7.2	47.7	28.9 ± 1.3	1/87
MCI	29	77.5 ± 7.1	51.7	27.0 ± 2.0	0/29
AD	28	73.6 ± 8.1	35.7	21.2 ± 5.6	0/28
CADDementia validation					
All	30	65.2 ± 7.0	43.3		0/30
NC	12	62.3 ± 6.3	25.0		0/12
MCI	9	68.0 ± 8.5	44.4		0/9
AD	9	66.1 ± 5.2	66.7		0/9
CADDementia test					
All	354	65.1 ± 7.8	60.2		0/354

[a] MMSE was not available for the CADDementia data.

The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and non-profit organizations, as a $60 million, 5-year, public-private partnership. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as to lessen the time and cost of clinical trials. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. For up-to-date information, see www.adni-info.org.

2.2.2. AIBL imaging arm

This dataset comprised the AIBL baseline imaging arm available via the ADNI database. The 3 T T1-weighted baseline structural MRI images were downloaded between September 27, 2013 and September 30, 2013. The data was collected by the AIBL study group. The AIBL study methodology has been reported previously (Ellis et al., 2009).

2.2.3. ADNI HHP subset

This dataset comprised a subset of 40 manual hippocampus segmentations from the Harmonized Hippocampal Protocol (HHP) (Boccardi et al., 2015) as well as the associated 1.5 T ADNI MRI scans reoriented along the anterior commissure (AC) – posterior commissure (PC) line using a six degrees of freedom linear transformation. We used the "preliminary release" version of the labels.

2.2.4. Training dataset (ADNI + AIBL)

The standardized ADNI dataset and the AIBL imaging arm were joined into one single training set for training of our method. This was possible because AIBL adopted the MRI protocol of ADNI, and because the neuropsychological tests in AIBL were designed to permit comparison and pooling with ADNI (Ellis et al., 2010). We denote the combined dataset ADNI + AIBL.

2.2.5. CADDementia validation and test set

The challenge used 3 T T1-weighted scans collected from the following three sites; Erasmus MC, Rotterdam, The Netherlands (174 scans); VU Medical Center, Amsterdam, The Netherlands (180 scans); and University of Porto/Hospital de São João, Porto, Portugal (30 scans). Two versions of the MRI data were available; raw data, and a bias field corrected and skull stripped version of the data. We used the raw data and applied the same bias field correction method to all MRI scans from ADNI, AIBL and CADDementia.

2.3. Algorithm

The algorithm was based on a number of individual MRI imaging biomarkers (Table 2); FreeSurfer cortical thickness measurements, FreeSurfer volumetric measurements, and hippocampal volume, shape and texture computed using special purpose methods. These biomarkers were z-score transformed within each group dependent on age, and entered as features to an LDA classifier as illustrated in Fig. 1.

Part of the FreeSurfer pipeline conforms the MRI scans to $1 \times 1 \times 1\text{mm}^3$ resolution and corrects for bias field using the non-parametric non-uniform intensity normalization (N3) algorithm (Sled et al., 1998). The hippocampal shape and hippocampal texture methods used this representation of the MRI data as well. The special purpose hippocampal volume method utilized the additional anonical atlas-based intensity normalization step in FreeSurfer. Hippocampal shape and texture were computed based on the FreeSurfer segmentation of the hippocampus because this is an established method.

Throughout the presentation, we represent the $N_{\text{class}} = 3$ diagnostic groups as $\omega_1 = \text{NC}$, $\omega_2 = \text{MCI}$, and $\omega_3 = \text{AD}$.

2.3.1. FreeSurfer volumetry

Sub-cortical, whole brain, and ventricular volumetric measurements were computed using cross-sectional FreeSurfer (Fischl et al., 2002). We used version 5.1.0 of FreeSurfer with default parameters. Bilateral ROIs were joined. All 7 volumetric measurements were normalized for head size by dividing by the intra-cranial volume (ICV)

Table 2
Overview of individual MRI biomarkers.

MRI biomarker	ROI segmentation	Training dataset
Cortical thickness		
Frontal lobe	FreeSurfer	No training
Occipital lobe	FreeSurfer	No training
Parietal lobe	FreeSurfer	No training
Temporal lobe	FreeSurfer	No training
Cingulate cortex	FreeSurfer	No training
Volumetry		
Amygdala	FreeSurfer	No training
Caudate nucleus	FreeSurfer	No training
Hippocampus	FreeSurfer	No training
Pallidum	FreeSurfer	No training
Putamen	FreeSurfer	No training
Ventricular	FreeSurfer	No training
Whole brain	FreeSurfer	No training
Special purpose hippocampus		
NL patch	NL patch	HHP
Left hipppocampus shape	FreeSurfer	ADNI + AIBL
Right hipppocampus shape	FreeSurfer	ADNI + AIBL
Hippocampal texture	FreeSurfer	ADNI + AIBL

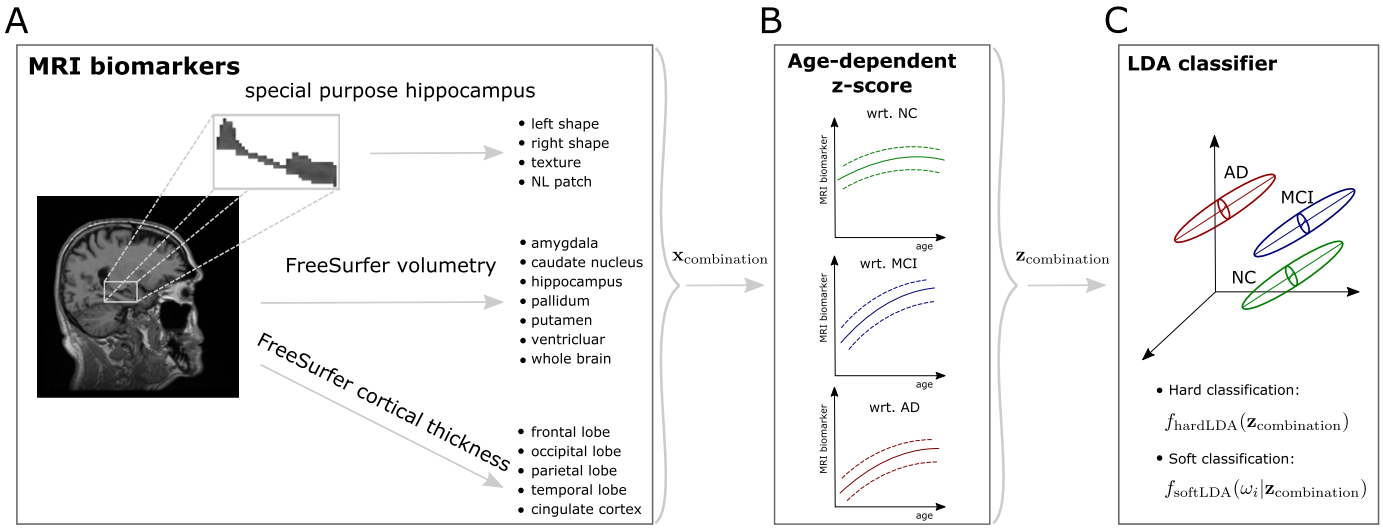

Fig. 1. Sketch illustrating the algorithm. (A) A range of individual structural MRI biomarkers capturing different aspects of the scan are extracted. (B) The individual MRI biomarkers are z-score normalized dependent on the age of the subject with respect to each diagnostic group. (C) The z-score normalized MRI biomarkers are entered to a 3-class LDA classifier to produce the combination biomarker score. A hard classification was obtained using $f_{\text{hardLDA}}(\mathbf{z}_{\text{combination}})$, and a soft classification was obtained using $f_{\text{softLDA}}(\omega_i|\mathbf{z}_{\text{combination}})$.

also computed during the cross-sectional FreeSurfer pipeline. The resulting feature vector looked as follows:

$$\mathbf{x}_{\text{volumetry}} = \begin{bmatrix} (|\text{left amygdala}| + |\text{right amygdala}|)/\text{ICV} \\ (|\text{left caudate nucleus}| + |\text{right caudate nucleus}|)/\text{ICV} \\ (|\text{left hippocampus}| + |\text{right hippocampus}|)/\text{ICV} \\ (|\text{left pallidum}| + |\text{right pallidum}|)/\text{ICV} \\ (|\text{left putamen}| + |\text{right putamen}|)/\text{ICV} \\ (|\text{ventricular}|)/\text{ICV} \\ (|\text{whole brain}|)/\text{ICV} \end{bmatrix}, \tag{1}$$

where $|\cdot|$ denotes volume of an ROI. Note that $|\cdot|$ is computed by FreeSurfer and that partial volume effects are accounted for in this computation.

2.3.2. FreeSurfer cortical thickness

Cortical thickness measurements were computed using cross-sectional FreeSurfer (version 5.1.0, default parameters) (Fischl and Dale, 2000). We used the Desikan-Killiany atlas that parcelates the entire cortex into 68 regions for each hemisphere. In order to keep the dimensionality of the feature vector low, the regions were joined into the four lobes and the cingulate cortex according to the specifications on the FreeSurfer website.[2] Left and right hemispheres were further joined, resulting in a total of 5 cortical thickness measurements. We did not normalize cortical thickness measurements for head size (i.e., ICV) (Westman et al., 2013). The cortical thickness feature vector looked as follows:

$$\mathbf{x}_{\text{cortical thickness}} = \begin{bmatrix} d(\text{left frontal lobe}) + d(\text{right frontal lobe}) \\ d(\text{left occipital lobe}) + d(\text{right occipital lobe}) \\ d(\text{left parietal lobe}) + d(\text{right parietal lobe}) \\ d(\text{left temporal lobe}) + d(\text{right temporal lobe}) \\ d(\text{left cingulate cortex}) + d(\text{right cingulate cortex}) \end{bmatrix}, \tag{2}$$

where $d(\cdot)$ computes the average distance between the gray/white matter boundary and the pial surface of an ROI. Note that d is computed by FreeSurfer.

2.3.3. Multi-atlas, patch-based hippocampal volume

In addition to the FreeSurfer estimate of the hippocampal volume, we also computed the hippocampal volume using a special purpose algorithm. This was motivated by the fact that hippocampal volume is the most widely used MRI biomarker of AD (Jack et al., 2011b), and since FreeSurfer segments many ROIs simultaneously in one objective function, it is not specific to a certain ROI. The left and right hippocampus were segmented separately using our own in-house implementation of the multi-atlas, non-local patch-based segmentation with expert priors technique (Coupé et al., 2011). Manual delineations from HHP (Boccardi et al., 2015), representing state-of-the-art in manual hippocampus segmentation, were used as expert priors (Anker, 2014; Anker et al., 2014). The method has previously demonstrated a better atrophy-based AD diagnostic performance than static FreeSurfer (Anker et al., 2014), and have in recent comparison studies performed by other research groups demonstrated better correspondence with manual segmentations than FreeSurfer (Manjón and Coupé, 2016; Næss-Schmidt et al., 2016).

The atlas comprised the 40 segmentations in the HHP subset and their corresponding MRI scans, both transformed to a common atlas space using affine registration. When segmenting a new test MRI scan, the N_{atlas} closest atlases were selected from the candidate set of 40 HHP segmentations by registering each HHP MRI scan in the atlas set to the test MRI using an affine registration and computing the sum of squared differences between intensities inside the FreeSurfer skull-stripped area. The N_{atlas} corresponding manual hippocampus delineations were used in the subsequent computations.

Following Coupé et al. (2011), the left and right hippocampus were segmented separately using the following steps. In order to avoid unnecessary computations, only voxels inside a coarse mask obtained by the union of the manual delineations in the atlas were segmented in the test MRI scan. The segmented label $\hat{y}_i$ of a voxel $\mathbf{x}_i$ in the test MRI scan was obtained by thresholding a weighted label

[2] http://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation (accessed 2016.10.20).

fusion of all labeled samples within the search volume from the N_{atlas} selected subjects

$$\hat{y}_i = \begin{cases} 1 \text{ if } \dfrac{\sum_{s=1}^{N_{\text{atlas}}} \sum_{j \in V} w(\mathbf{x}_i, \mathbf{x}_{s,j}) y_{s,j}}{\sum_{s=1}^{N_{\text{atlas}}} \sum_{j \in V} w(\mathbf{x}_i, \mathbf{x}_{s,j})} \geq 0.5 \\ 0 \text{ otherwise} \end{cases}, \tag{3}$$

where $\mathbf{x}_{s,j}$ is the voxel at position j in atlas s with corresponding HHP segmentation label $y_{s,j} \in \{0, 1\}$ encoded according to whether voxel $\mathbf{x}_{s,j}$ is background or hippocampus, V is the search volume, and the label weights according to the similarity of the intensities in the two patches being compared is

$$w(\mathbf{x}, \mathbf{x}_{j,s}) = \begin{cases} \exp\left(\frac{-||p(\mathbf{x}), p(\mathbf{x}_{s,j})||_2^2}{n^3 h(\mathbf{x})}\right) & \text{if } \tau(\mathbf{x}, \mathbf{x}_{j,s}) > 0.95 \\ 0 & \text{otherwise} \end{cases},$$

$$\tau(\mathbf{x}, \mathbf{x}_{j,s}) = \frac{2\mu_{p(\mathbf{x})}\mu_{p(\mathbf{x}_{s,j})}}{\mu_{p(\mathbf{x})}^2 + \mu_{p(\mathbf{x}_{s,j})}^2} \times \frac{2\sigma_{p(\mathbf{x})}\sigma_{p(\mathbf{x}_{s,j})}}{\sigma_{p(\mathbf{x})}^2 + \sigma_{p(\mathbf{x}_{s,j})}^2}. \tag{4}$$

Here $p(\mathbf{x})$ extracts the n^3-dimensional vector of concatenated intensities in the $n \times n \times n$ patch centered on $\mathbf{x}$, h is a smoothing parameter that locally adapts the similarity measure according to the distance of the closest atlas patch, and $\mu_{f(\mathbf{x})}$ and $\sigma_{f(\mathbf{x})}$ are the mean and standard deviation of the intensities in patch $f(\mathbf{x})$, respectively. In Eqs. (3) and (4), the thresholds were selected according to Coupé et al. (2011). The threshold on τ serves as pre-selection of patches to reduce computational time. That is, dissimilar atlas patches are disregarded in the label fusion where similarity, τ, is computed as the product of simple measures of luminance and contrast difference. As proposed by Coupé et al. (2012), h was estimated according to

$$h(\mathbf{x}) = 0.25 \min_{\mathbf{x}_{s,j}} ||p(\mathbf{x}) - p(\mathbf{x}_{s,j})||_2 + \epsilon,$$

where ϵ is a small constant to avoid division by zero when the test patch is contained in the atlas.

The following parameter values were determined as the ones maximizing the leave-one-out estimation of Dice's coefficient on a subset of 15 HHP segmentations:

- The number of atlases considered in initial atlas selection (N_{atlas}) was set to 9.
- A patch size (n) of 5 mm was used.
- A search volume (V) size of 9 × 9 × 9 mm was used.

The final non-local patch (NL patch) feature was

$$\mathbf{x}_{\text{NL patch}} = \frac{|\text{left hippocampus}| + |\text{right hippocampus}|}{\text{ICV}}, \tag{5}$$

where ICV was estimated from the same MRI scan using FreeSurfer.

2.3.4. Hippocampal shape

Two hippocampal shape scores, one for the left and one for the right hippocampus, were computed as well. Each hippocampus was represented by a shape descriptor that was subsequently classified using a naïve Bayes classifier trained explicitly for left or right hipppocampus.

In a spirit similar to Achterberg et al. (2014), the left and right hippocampus shape descriptors were computed using the following four steps:

1. A random subject was selected from the ADNI dataset as template and represented by 30 landmarks uniformly distributed across the surface of its FreeSurfer hippocampus segmentation.
2. Each hippocampus was aligned to the template hippocampus using iterative closest point between the two respective FreeSurfer segmentations treated as point clouds. Subsequently, the pre-defined template surface landmarks were mapped to the aligned hippocampus by selecting its FreeSurfer segmentation surface points closest to the template surface points. As a result, each hippocampus was now represented by 30 surface landmarks, $\mathbf{x} = [x_1, y_1, z_1, \ldots, x_{30}, y_{30}, z_{30}]^T$.
3. The set of hippocampi represented as surface landmarks, $\{\mathbf{x}\}$, were all aligned using generalized Procrustes alignment (Gower, 1975), producing a set of aligned shapes $\{\hat{\mathbf{x}}\}$.
4. Finally, principal component analysis (PCA) was applied to the set of aligned hippocampus landmarks $\{\hat{\mathbf{x}}\}$, and the $N_{\text{component}}$ components explaining 90% of the variance were retained, resulting in the final per-hippocampus shape descriptor $\tilde{\mathbf{x}} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{N_{\text{component}}}]^T$.

This representation was used as features in a 2-class Gaussian naïve Bayes classifier computing the posterior probability of AD

$$f_{\text{naive Bayes}}(\text{AD}|\tilde{\mathbf{x}}) = \frac{\exp g_3(\tilde{\mathbf{x}}))}{\exp(g_1(\tilde{\mathbf{x}})) + \exp(g_3(\tilde{\mathbf{x}}))}, \tag{6}$$

where $g_1(\tilde{\mathbf{x}})$ and $g_3(\tilde{\mathbf{x}})$ are defined as

$$g_i(\tilde{\mathbf{x}}) = \sum_{j=1}^{N_{\text{component}}} \left(-\log\left(\sqrt{2\pi}\sigma_{\omega_i j}\right) - \frac{(\tilde{x}_j - \mu_{\omega_i j})^2}{2\sigma_{\omega_i j}^2}\right) + \log P(\omega_i)$$

for $i \in \{1, 3\}$, where $\mu_{\omega_i j}$ and $\sigma_{\omega_i j}$ are the class-conditional mean and standard deviation for the jth component, and $P(\omega_i)$ is the class prior.

The feature extraction was performed on all data simultaneously, i.e., on the combination of the training set and the test set. Subsequently, only NC and AD observations from the training set were used for training of the naïve Bayes classifier. The trained classifier was finally applied to score the test data. From a machine learning perspective, that means we make use of transductive inference (Vapnik, 1998, chap. 8). The class priors in (6) were set to the class frequencies in the training set. The whole procedure was computed for the left and the right hippocampus separately, resulting in two shape scores. The FreeSurfer hippocampus segmentation was used to defined the left and right hippocampus ROI in each MRI scan. The shape feature vector looked as follows:

$$\mathbf{x}_{\text{shape}} = \begin{bmatrix} f_{\text{naiveBayes}}(\mathbf{x}_{\text{leftShapeDescriptor}}) \\ f_{\text{naiveBayes}}(\mathbf{x}_{\text{rightShapeDescriptor}}) \end{bmatrix}, \tag{7}$$

where $\tilde{\mathbf{x}}$ was computed for the left and right hippocampus separately to produce $\mathbf{x}_{\text{leftShapeDescriptor}}$ and $\mathbf{x}_{\text{rightShapeDescriptor}}$.

2.3.5. Hippocampal texture

A hippocampal texture score was computed using a texture descriptor that has previously been successfully applied for quantification of chronic obstructive pulmonary disease in computed tomography (Sørensen et al., 2012) in combination with a support vector machine (SVM) (Cortes and Vapnik, 1995) with a radial Gaussian kernel. This specific MRI biomarker has previously shown good AD diagnostic results, captures independent information from volume, and has demonstrated capabilities of earlier AD detection than volume (Sørensen et al., 2016). Texture therefore complements volume well.

The texture descriptor comprised marginal filter response histograms of a 3-dimensional, rotation-invariant, multi-scale, Gaussian derivative-based filter bank (Lindeberg, 2008). The histograms were computed using filter responses from both hippocampi collectively. These histograms could capture different micro-structural properties

within the hippocampal tissue, such as the amount of steep intensity transitions and "blob"-like structures. The descriptor was adapted to our problem and therefore deviated from Sørensen et al. (2012) in the following four ways:

1. The Gaussian filter was excluded in order to be invariant to the lack of a standard image intensity scale in MRI (Nyúl and Udupa, 1999). This exclusion left the following seven base filters measuring different aspects of the local image structure: the three eigenvalues of the Hessian matrix, gradient magnitude, the Laplacian of the Gaussian, Gaussian curvature, and the Frobenius norm of the Hessian matrix. All these filters are based on intensity derivatives, and the method is therefore invariant to locally constant intensity offsets (e.g., caused by imperfections in the N3 bias field correction).
2. The following scales were used: 0.6, 0.85, 1.2, and 1.7 mm. The upper end of the scale range was determined by visual inspection of Gaussian smoothed images. The structures in the hippocampus visually vanished at scales exceeding 1.7 mm.
3. Derivatives at the different scales were computed by convolution with the corresponding derivative filter instead of convolution with a Gaussian followed by finite differencing for improved numerical accuracy.
4. Based on the size of the smallest morphologically cleaned bilateral hippocampal segmentation in the ADNI dataset, we quantized the filter responses into nine histogram bins. The descriptor was applied to the conformed MRI scans, and since the FreeSurfer conformation and the filtering are both linear processes, their combination is mathematically equivalent to one linear process.

In the following, $I_{x,\sigma}$ and $I_{xx,\sigma}$ denote the partial first order and second order derivative, respectively, of MRI image I w.r.t. x at scale σ, and $\mathbf{x} = [x, y, z]^T$ is a voxel. Partial derivatives of the image at the different scales in the multi-scale representation of the image were in all cases obtained using Gaussian derivatives (Lindeberg, 2008). For example, $I_{x,\sigma} = I * G_{x,\sigma}$ where $*$ denotes convolution and $G_{x,\sigma}$ is the partial first-order derivative of the Gaussian function,

$$G(\mathbf{x};\sigma) = \frac{1}{((2\pi)^{(1/2)}\sigma)^3}\exp\left(-\frac{||\mathbf{x}||_2^2}{2\sigma^2}\right),$$

w.r.t. x at scale σ. This way of defining derivatives in scale-space makes the computation of image derivatives well-posed (Lindeberg, 2008). The seven base filters in the filter bank of the texture descriptor were the following: the three eigenvalues of the Hessian matrix

$$\lambda_i(\mathbf{x};\sigma), \quad i \in \{1,2,3\}, \quad |\lambda_1(\mathbf{x};\sigma)| \geq |\lambda_2(\mathbf{x};\sigma)| \geq |\lambda_3(\mathbf{x};\sigma)|,$$

where

$$H(\mathbf{x};\sigma) = \begin{bmatrix} I_{xx,\sigma} & I_{xy,\sigma} & I_{xz,\sigma} \\ I_{xy,\sigma} & I_{yy,\sigma} & I_{yz,\sigma} \\ I_{xz,\sigma} & I_{yz,\sigma} & I_{zz,\sigma} \end{bmatrix};$$

gradient magnitude

$$||\nabla G(\mathbf{x};\sigma)||_2 = \sqrt{I_{x,\sigma}^2 + I_{y,\sigma}^2 + I_{z,\sigma}^2};$$

Laplacian of the Gaussian

$$\nabla^2 G(\mathbf{x};\sigma) = \lambda_1(\mathbf{x};\sigma) + \lambda_2(\mathbf{x};\sigma) + \lambda_3(\mathbf{x};\sigma);$$

Gaussian curvature

$$K(\mathbf{x};\sigma) = \lambda_1(\mathbf{x};\sigma)\lambda_2(\mathbf{x};\sigma)\lambda_3(\mathbf{x};\sigma);$$

and the Frobenius norm of the Hessian

$$||H(\mathbf{x};\sigma)||_F = \sqrt{\lambda_1(\mathbf{x};\sigma)^2 + \lambda_2(\mathbf{x};\sigma)^2 + \lambda_3(\mathbf{x};\sigma)^2}.$$

The histogram of filter responses inside the hippocampus in a filtered MRI image was estimated as

$$h_f(i,I) = \frac{\sum_{\mathbf{x}\in S}\delta_i(I_f(\mathbf{x}))}{|S|}, \quad i = 1\ldots 9,$$

where S is the joined left and right hippocampus segmentation computed from I using the static FreeSurfer pipeline, I_f is I filtered using filter f, and $\delta_i(\cdot)$ is an indicator function defined as

$$\delta_i(I_f(\mathbf{x})) = \begin{cases} 1 & \text{if } \Delta_i < I_f(\mathbf{x}) < \Delta_{i+1} \\ 0 & \text{otherwise} \end{cases},$$

where Δ_i is a histogram bin edge. The edges were determined using adaptive binning (Ojala et al., 1996); Δ_i is the $(i-1)/N_{\text{bin}} \times 100$th percentile filter response in the training set where N_{bin} is the number of histogram bins. In this work, morphologically post-processed bilateral FreeSurfer hippocampus segmentations were used to define the ROI, and the number of histogram bins was determined as $N_{\text{bin}} = \lfloor\sqrt[3]{731}\rfloor = 9$ in the experiments. The smallest morphologically cleaned bilateral FreeSurfer hippocampal segmentation in the ADNI dataset, which contained 731 voxels, determined this number.

The final hippocampal texture descriptor was obtained by concatenating the filter response histograms;

$$\mathbf{x}_{\text{textureDescriptor}} = [h_{\lambda_1(x;0.6)}, \ldots, h_{||H(x;0.6)||_F}, \ldots, h_{||H(x;1.7)||_F}]^T.$$

The concatenated histograms were used as input features x to a support vector machine (SVM) (Cortes and Vapnik, 1995) trained using the NC and AD observations from the training set. The SVM discriminant function

$$f_{\text{SVM}}(x) = \sum_{i=1}^{N_{\text{train}}} a_i k(x_i, x) + b$$

was used for texture scoring. It is determined by solving

$$\underset{a\in\mathbb{R}^{N_{\text{train}}}, b\in\mathbb{R}}{\text{minimize}} \sum_{i=1}^{N_{\text{train}}} L_{\text{hinge}}(y_i, f_{\text{SVM}}(x_i)) + \frac{1}{2C}\sum_{i=1,j=1}^{N_{\text{train}}} a_i a_j k(x_i, x_j)$$

where $x_i \in \mathcal{X}$ $(i = 1, \ldots, N_{\text{train}})$ is a training pattern and $y_i \in \{-1, 1\}$ encodes the class label according to whether pattern i is labeled NC or AD, the loss function is defined as $L_{\text{hinge}}(y, \hat{y}) = \max(0, 1 - y\hat{y})$, and $k(\cdot,\cdot)$ is the radial Gaussian kernel,

$$k(x_i, x) = \exp(-\gamma||x_i - x||_2^2).$$

The regularization parameter $C > 0$ and the kernel parameter $\gamma > 0$ were selected using grid search. We considered the following hyperparameter values in the grid: $\log(C) \in \{0, 1, \ldots, 10\}$ and $\log(\gamma) \in \{\log(\gamma_{\text{Jaakkola}}) - 4 + 1/3i\}_{i=0,1,\ldots,24}$, where

$$\gamma_{\text{Jaakkola}} = \text{median}\{\min\{||x_i - x_j||_2 \,|\, (x_j, y_j) \in S \wedge y_i \neq y_j\} | (x_i, y_i) \in S\}$$

provided an initial guess for γ around which the grid search was centered (Jaakkola et al., 1999). The performance of each parameter combination (C, γ) was estimated using 20-fold cross-validation splitting the available training data stratified by class label. The parameter combination that resulted in the lowest cross-validation AUC was selected, and the SVM was finally trained on the complete training set using these parameters.

The final hippocampal texture score was obtained by applying the trained SVM to the textural description of an observation

$$\mathbf{x}_{\text{texture}} = f_{\text{SVM}}(\mathbf{x}_{\text{textureDescriptor}}). \quad (8)$$

2.3.6. Age-dependent z-score transformation and classification

The individual MRI biomarker feature vectors, Eqs. (1), (2), (5), (6) and (7), were concatenated into one 16-dimensional combined feature vector

$$\mathbf{x}_{\text{combination}} = [\mathbf{x}_{\text{volumetric}}, \mathbf{x}_{\text{cortical thickness}}, \mathbf{x}_{\text{NL patch}}, \mathbf{x}_{\text{shape}}, \mathbf{x}_{\text{texture}}]^T.$$

Each entry x in $\mathbf{x}_{\text{combination}}$ was normalized for age using a z-score transformed dependent on the age of the subject according to $z = (x - \mu_{\text{age}})/\sigma_{\text{age}}$. The age-dependent weighted mean, μ_{age}, and the age-dependent weighted standard deviation, σ_{age}, of the biomarker used in the transformation were estimated from the training set using an adaptive width Gaussian interpolation kernel centered on the respective age. The age-dependent z-score transformation was applied for each diagnostic group, resulting in a 48-dimensional feature vector $\mathbf{z}_{\text{combination}}$ that was used as the final representation in the algorithm.

The combined and z-score transformed MRI biomarkers were classified using a 3-class LDA classifier. Either as a hard classification

$$f_{\text{hardLDA}}(\mathbf{z}_{\text{combination}}) = \underset{i=1,\ldots,N_{\text{class}}}{\text{argmax}} \; g_i(\mathbf{z}_{\text{combination}})$$

or as a posterior probability computed using the softmax function

$$f_{\text{softLDA}}(\omega_i|\mathbf{z}_{\text{combination}}) = \frac{\exp(g_i(\mathbf{z}_{\text{combination}}))}{\sum_{j=1}^{N_{\text{class}}} \exp(g_j(\mathbf{z}_{\text{combination}}))}.$$

In both cases, $g_i(\cdot)$ is the LDA discriminant function

$$g_i(\mathbf{z}_{\text{combination}}) = \mathbf{z}_{\text{combination}}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{\omega_i} - \frac{1}{2} \boldsymbol{\mu}_{\omega_i}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{\omega_i} + \log P(\omega_i)$$

where $\boldsymbol{\mu}_{\omega_i}$ is the class-conditional mean and $\boldsymbol{\Sigma}$ is the pooled covariance matrix. LDA training and classification was performed using the open source C++ machine learning library Shark (Igel et al., 2008). The class priors in were set to the class frequencies in the training set.

3. Results and discussion

The CADDementia data was scored using the LDA classifier trained on all ADNI + AIBL data. The ADNI + AIBL dataset was scored using 10-fold cross-validation stratified by cohort (ADNI, AIBL) and diagnostic group (NC, MCI, AD). The same folds were used to obtain individual hippocampal shape and hippocampal texture scores, and to obtain scores from the combination LDA. When scoring CADDementia data, hippocampal shape features were extracted from the combination of ADNI + AIBL and either the CADDementia validation set (the number of components, $N_{\text{component}}$, retained in the PCA were 50 for the left hippocampus and 48 for the right hippocampus) or the CADDementia test set (51 PCA components for the left hippocampus and 49 for the right hippocampus), dependent on which dataset was to be scored, and in the 10-fold cross-validation procedure used to score the ADNI + AIBL dataset, hippocampal shape features were extracted using the entire ADNI + AIBL dataset (50 PCA components for the left hippocampus and 48 for the right hippocampus). Because the shape features were extracted in this way, we perform transductive inference.

FreeSurfer failed to process three scans in the CADDementia test set. This was handled by imputing features. For two cases, the FreeSurfer pipeline had reached past the sub-cortical segmentation step, but not far enough to produce stats-files with partial volume corrected volume estimates, nor cortical thickness measurements. For these cases, we assigned the mean of the class with the most similar texture score because we could still compute this. The last cased failed before sub-cortical segmentation. Here we performed a visual inspection and assigned the mean of the NC class.[3]

3.1. Diagnostic results

Diagnostic measures for the CADDementia validation set and ADNI + AIBL were obtained using the Python scripts supplied by the CADDementia team on our scores, and the CADDementia test set performance was obtained from the challenge itself.

The combination biomarker achieved a CA of 63.0% and a total AUC of 78.5 on the CADDementia challenge test set (Table 3). This was comparable to the CA and total AUC estimated using 10-fold cross-validation on ADNI + AIBL. The performance on the CADDementia validation set was substantially better with a CA of 73.3% and a total AUC of 83.2. We attribute this discrepancy to the rather small size of the CADDementia validation set (30 observations). Better performance on the CADDementia validation set was observed across all participating teams (Bron et al., 2015). The TFPs and the confusion matrices in Table 4 reveals that on a diagnostic group level, the results between the CADDementia test set and ADNI+AIBL were less comparable with absolute true positive fraction differences in the range 17.9% to 29.1%. The discrepancy could be explained by cohort differences. For example, the CADDementia subjects are on average 10 years younger than ADNI and AIBL subjects (Table 1). However, we would not expect age to be the primary cause because of the age-dependent z-score transformation of the individual MRI biomarkers prior to entering the LDA. Another possible cause is difference in MRI field strength, the CADDementia data is 3 T whereas ADNI is 1.5 T. However, the addition of the 144 3 T AIBL scans to the training set should to some degree account for this. Moreover, we would expect field strength to play a smaller role because this would imply a shift in features across the entire dataset, not within specific diagnostic groups. Another potential cause is differences in the criteria for the diagnostic groups, i.e., how the labels are defined. Approximately half the CADDementia NC group are controls with subjective complaints (Bron et al., 2015) as opposed to only 20% in ADNI + AIBL. ADNI explicitly defines NC to be controls with no memory complaints (Petersen et al., 2010), and by design, approximately 50% of controls in AIBL had subjective memory complaints (Ellis et al., 2010). ADNI and AIBL include only mild AD patients at baseline (Petersen et al., 2010;Ellis et al., 2010) as opposed to CADDementia that does not restrict the severity (Bron et al., 2015). The AD patients in the training dataset had mean mini-mental state examination (MMSE) scores of 23.2 ± 1.9 and 21.2 ± 5.6, for the ADNI and AIBL parts, respectively. In comparison, the AD patients from the VU Medical Center, which amounts to approximately half of the AD patients in the CADDementia dataset, had a mean MMSE score of 20 ± 4.6 (Binnewijzend et al., 2013).

[3] Another team participating in the CADDementia challenge also reported problems in FreeSurfer processing of 3 cases using FreeSurfer v5.1 (Wachinger et al., 2014).

Table 3
Performance measures.

	CA	True positive fraction			AUC			
		NC	MCI	AD	All	NC	MCI	AD
ADNI + AIBL[a]	62.7	79.0	57.8	40.3	78.1	86.1	68.3	81.8
CADDementia validation	73.3	91.7	44.4	77.8	83.2	86.6	68.3	95.8
CADDementia test[b]	63.0	96.9	28.7	61.2	78.8	86.3	63.1	87.5

[a] 10-Fold cross-validation stratified by cohort and diagnostic group.
[b] Results from Bron et al. (2015).

In comparison to the results of the other participating teams in the CADDementia challenge, the 63.0% CA was significantly better than the methods not in top-six according to a McNemar's test on the hard classification output (Bron et al., 2015). The total AUC of 78.5 was also the highest among all entries and only the top-five methods had AUCs within the lower AUC confidence interval of our method (Bron et al., 2015).

3.2. Feature contribution

In order to inspect the contribution of each feature in classification, we performed a feature selection experiment using sequential forward feature selection (SFS) (Jain et al., 2000). The objective function in the SFS procedure was the 20-fold cross-validation CA of an LDA classifier applied to the training set. We used the training set folds from the ADNI + AIBL 10FCV dataset and computed the average CA curve as a function of the number of selected features across the 10 folds (Fig. 2A). The curve converged at 10 features, and this was the number of features retained in the subsequent analysis. The frequency of selection when keeping the first 10 features in each of the 10 folds was computed (Fig. 2B). In this analysis, a feature could be selected a maximum of 30 times (3 different z-score representations across 10 folds). The most frequently selected features were as follows (reported as % of possible): FreeSurfer hippocampal volume (66.7%), FreeSurfer ventricular volume (53.3%), hippocampal texture (50.0%), FreeSurfer parietal lobe thickness (46.7%), right hippocampal shape (23.3%), FreeSurfer occipital lobe thickness (20.0%), and FreeSurfer cingulate cortex thickness (20.0%). In addition, we analyzed how early a feature was selected (indicated by color-coding in Fig. 2B). Hippocampal texture was consistently selected in the first iteration of the SFS procedure in all 10 folds, FreeSurfer hippocampal volume was selected in the second iteration in 9 folds and in the third iteration in 1 fold, and FreeSurfer ventricular volume was selected in 1 fold in the second iteration and in 6 folds in the third iteration. The following feature subsets occurred most frequently in the 10 folds: {hippocampal texture, FreeSurfer hippocampal volume, FreeSurfer ventricular volume} (10 folds), {hippocampal texture, FreeSurfer hippocampal volume, FreeSurfer ventricular volume, FreeSurfer parietal lobe thickness} (9 folds), {hippocampal texture, FreeSurfer hippocampal volume, FreeSurfer ventricular volume, FreeSurfer occipital lobe thickness} (6 folds), and {hippocampal texture, right hippocampus shape, FreeSurfer hippocampal volume, FreeSurfer ventricular volume} (5 folds). The reduced feature representation (i.e., using the first 10 features as selected by SFS on the training set in each fold of the 10-fold cross-validation procedure) resulted in a 62.6% CA on the ADNI + AIBL dataset. This was similar to the 62.7% CA obtained when using the full feature representation.

To gain insight into which features contribute to discrimination between two specific diagnostic groups, we repeated the SFS for each pairwise scenario (NC vs. MCI, NC vs. AD, MCI vs. AD) using a 2-class LDA as objective function. When considering NC vs. MCI, the picture was similar to the 3-class scenario; hippocampal texture was consistently selected as the first feature in all 10 folds, and FreeSurfer hippocampal volume (50.0%) and ventricular volume (50.0%) were alongside hippocampal texture (56.7%) the most frequently selected features. However, the selection frequency for FreeSurfer parietal lobe thickness dropped to 30.0%, whereas FreeSurfer occipital lobe thickness increased its frequency to 33.3%. The picture changed completely when considering MCI vs. AD. In this case, the FreeSurfer temporal lobe thickness was selected as the first feature in 8 out of 10 folds and FreeSurfer parietal lobe thickness was selected first in the 2 remaining folds, and the distribution of selection was more uniform with FreeSurfer amygdala volume (40.0%), FreeSurfer temporal lobe thickness (40.0%), NL patch (36.7%), and FreeSurfer occipital lobe thickness (33.3%) as the most frequently selected features. For NC vs. AD, the frequency of selection was the most peaked and a mix of previous tendencies was observed. Hippocampal texture was selected first in 9 out of 10 folds whereas FreeSurfer temporal lobe thickness was selected in the remaining fold. The most frequent features were as follows: FreeSurfer temporal lobe thickness (63.3%), FreeSurfer hippocampus volume (50.0%), and hippocampal texture (36.7%).

3.3. Feature relations

We inspected the relationship between the features entered to the LDA by computing their pair-wise Pearson correlation (Table 5). Prior to computing correlation, each feature was aggregated across the 3 per-group z-scores by computing the mean feature value. As expected, the most correlated features were the two hippocampal volume estimates from FreeSurfer and NL patch that had a correlation of $\rho = 0.9$. Other expected high correlations included the volume of the hippocampus and of the amygdala (Poulin et al., 2011) ($\rho = 0.8$ and $\rho = 0.7$, depending on hippocampal volume estimation method), and hippocampal volume vs. hippocampal texture ($\rho = -0.7$ and $\rho = -0.6$). The cortical thickness features were generally highly correlated, with one insignicifant correlation and 9 significant correlations ranging from $\rho = 0.5$ to $\rho = 0.8$ among the 10 pair-wise combinations. Finally, we note that the temporal lobe was the only cortical thickness feature that correlated with

Table 4
Confusion matrices. Rows are predicted class and columns are true class.

ADNI+AIBL[a]				CADDementia validation				CADDementia test[b]			
	NC	MCI	AD		NC	MCI	AD		NC	MCI	AD
NC	203	65	9	NC	11	3	0	NC	125	64	15
MCI	48	152	68	MCI	1	5	2	MCI	3	35	25
AD	6	46	52	AD	0	1	7	AD	1	23	63

[a] 10-fold cross-validation stratified by cohort and diagnostic group.
[b] results from Bron et al. (2015).

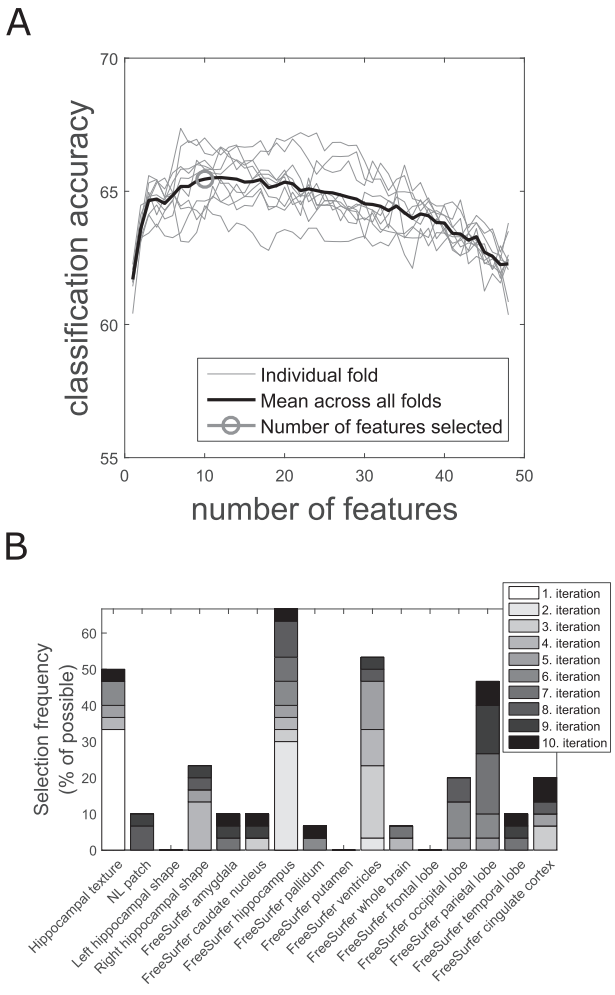

Fig. 2. Result of sequential forward feature selection (SFS). (A) Performance as a function of number of selected features. The thin gray lines correspond to each of the 10 training sets in the ADNI+AIBL 10FCV dataset, and the thick black line is the average performance across the 10 training folds. The average curve converges at 10 features indicated by the gray circle. This is the number of features that is used in the subsequent analysis. (B) Frequency of selection when the first 10 features are considered in the SFS. Each feature can be selected maximally 30 times (corresponding to the 3 different z-score versions across the 10 folds). The color-coding corresponds to how early in the SFS procedure a feature is selected, and it ranges from white (1. iteration) to black (10. and last iteration).

volumetric and special purpose hippocampus features, and that the highest correlation was with hippocampal texture ($\rho = -0.7$).

3.4. Sufficient training data?

The learning curves of the combination biomarker were plotted in order to determine whether the method would have benefited from more training data. In a spirit similar to Perlich et al. (2003), the curves were computed based on ADNI + AIBL as follows:

1. ADNI + AIBL is initially split randomly stratified by cohort and diagnostic group into a validation set $X_{\text{validation}}$ of size 49 and the remaining 600 observations are kept as training data pool $X_{\text{train-source}}$.
2. Repeat for $N = 50, 100, \ldots, 600$.
   (a) A training set X_{train} of current size N is sampled without replacement from $X_{\text{train-source}}$.
   (b) The combination LDA is trained on X_{train} and is subsequently applied to score both X_{train} and $X_{\text{validation}}$. The resulting classification accuracies are recorded.

The above procedure is repeated 100 times, and the mean and standard deviation of the resulting training curves produce the final training learning curve. The final validation learning curve was obtained in a similar manner.

The curves are converging and approach each other (Fig. 3A), suggesting that further increasing the amount of training data would only have a small effect on the performance. The learning task appears to be a "high bias" problem. However, one should also bear in mind that the Bayes optimal error for this problem is not zero. First, there is noise in the labels because the diagnosis is clinical and a definite diagnosis would only be possible post-mortem. CADDementia uses the NINCDS/ADRDA criteria for probable AD (McKhann et al., 1984) to define the AD group, and an average sensitivity of 81% across several studies at a specificity of 70% have been reported for these criteria when compared to neurological confirmation (Knopman et al., 2001). The MCI group is also based on a clinical diagnosis (Petersen, 2004), and MCI is in general a heterogeneous entity. Both properties may lead to label noise as well. Secondly, we expect there is a limit as to what structural MRI is capable of in isolation, especially when MCI is part of the problem. According to recent hypothetical models of AD biomarker dynamics, structural MRI is one of the biomarkers that are dynamic late during the disease process (Jack et al., 2013).

3.5. A more complex classifier?

One way to approach a high bias problem is to use a more flexible classifier, possibly in combination with more training data. In order to test this, we repeated the learning curve experiment with the following three non-linear classifiers:

- a radial Gaussian SVM. Regularization and kernel parameter were determined using grid search, and the performance of each parameter combination was estimated using 20-fold cross-validation on the training set in a particular fold;
- a random forest classifier. The split in each node of a tree in the random forest classifier was done on $\lfloor\sqrt{48}\rfloor = 6$ features (Hastie et al., 2009), and 500 trees were used in the ensemble;
- a k nearest neighbor (kNN) classifier. Euclidean distance and k chosen according to the usual square root rule (Alkoot and Kittler, 2002): $k = \lfloor\sqrt{N_{\text{train}}}\rfloor$ where N_{train} is the number of training set observations in a particular fold.

It was observed that the curves of the non-linear classifiers converged to a performance similar to the LDA, and that the training and validation curves met each other (Fig. 3B–D). Therefore, we conjecture that for the given features, we are close to the Bayes optimal solution.

3.6. General discussion

Structural MRI is among the core biomarkers of AD (Jack et al., 2013). It is considered a surrogate marker of neurodegeneration and has generally been shown to be sensitive relatively late during the course of the disease. However, the placement of structural MRI in the hypothetical model of AD biomarker dynamics put forward by Jack et al. (2013) is based on evidence from volumetric MRI studies. There are other MRI-based biomarkers that target subtler information, such as the hippocampal shape and hippocampal texture, both considered in this study. These MRI biomarkers have been shown to be predictive of dementia independent of volume (Achterberg et al., 2014; Sørensen et al., 2016), and there is reason to believe that combining volume with these markers would increase the range over which structural MRI is sensitive to the course of AD.

Table 5
Pair-wise Pearson correlation between features. Asterisk marks significance Bonferroni corrected across pair-wise comparisons ($p < 0.000008 = 0.001/((16^2 - 16)/2)$). Bold font marks significant correlations of at least 0.5.

	FreeSurfer cortical thickness					FreeSurfer volumetry							Special purpose hippocampus		
	FL	OL	PL	TL	CC	AM	CN	HI	PA	PU	VE	WB	NL patch	Shape (l)	Shape (r)
OL	**0.6***														
PL	**0.8***	**0.8***													
TL	**0.7***	**0.6***	**0.7***												
CC	**0.7***	0.4*	**0.5***	**0.6***											
AM	0.3*	0.4*	0.4*	**0.6***	0.3*										
CN	0.1	0.2*	0.2*	0.1	0.1	0.2*									
HI	0.4*	0.3*	0.4*	**0.6***	0.4*	**0.8***	0.1								
PA	0.2*	0.2*	0.2*	0.2*	0.1	0.3*	0.3*	0.4*							
PU	0.3*	0.4*	0.4*	0.4*	0.3*	**0.5***	**0.6***	**0.5***	**0.5***						
VE	−0.3*	−0.2	−0.2*	−0.4*	−0.3*	−0.3*	0.1	**−0.5***	−0.4*	−0.4*					
WB	0.2*	0.2*	0.2*	0.2*	0.2*	0.4*	0.3*	**0.5***	**0.5***	**0.5***	−0.3*				
NL patch	0.3*	0.2*	0.2*	**0.5***	0.3*	**0.7***	0.1	**0.9***	0.4*	0.4*	−0.4*	0.4*			
Shape (l)	−0.2*	−0.2*	−0.2*	**−0.5***	−0.1	−0.4*	0.0	**−0.5***	−0.2*	−0.2*	0.3*	−0.2*	**−0.5***		
Shape (r)	−0.1	0.0	−0.1	−0.1	0.0	0.0	0.0	−0.1	0.0	0.0	0.1	0.0	−0.1	0.1	
Texture	−0.4*	−0.4*	−0.4*	**−0.7***	−0.3*	**−0.6***	−0.1	**−0.7***	−0.2*	−0.3*	0.4*	−0.3*	**−0.6***	**0.5***	0.1

Abbreviations: FL, frontal lobe; OL, occipital lobe; PL, parietal lobe; TL, temporal lobe; CC, cingulate cortex; AM, amygdala; CN, caudate nucleus; HI, hippocampus; PA, pallidum; PU, putamen; VE, ventricular; WB, whole brain.

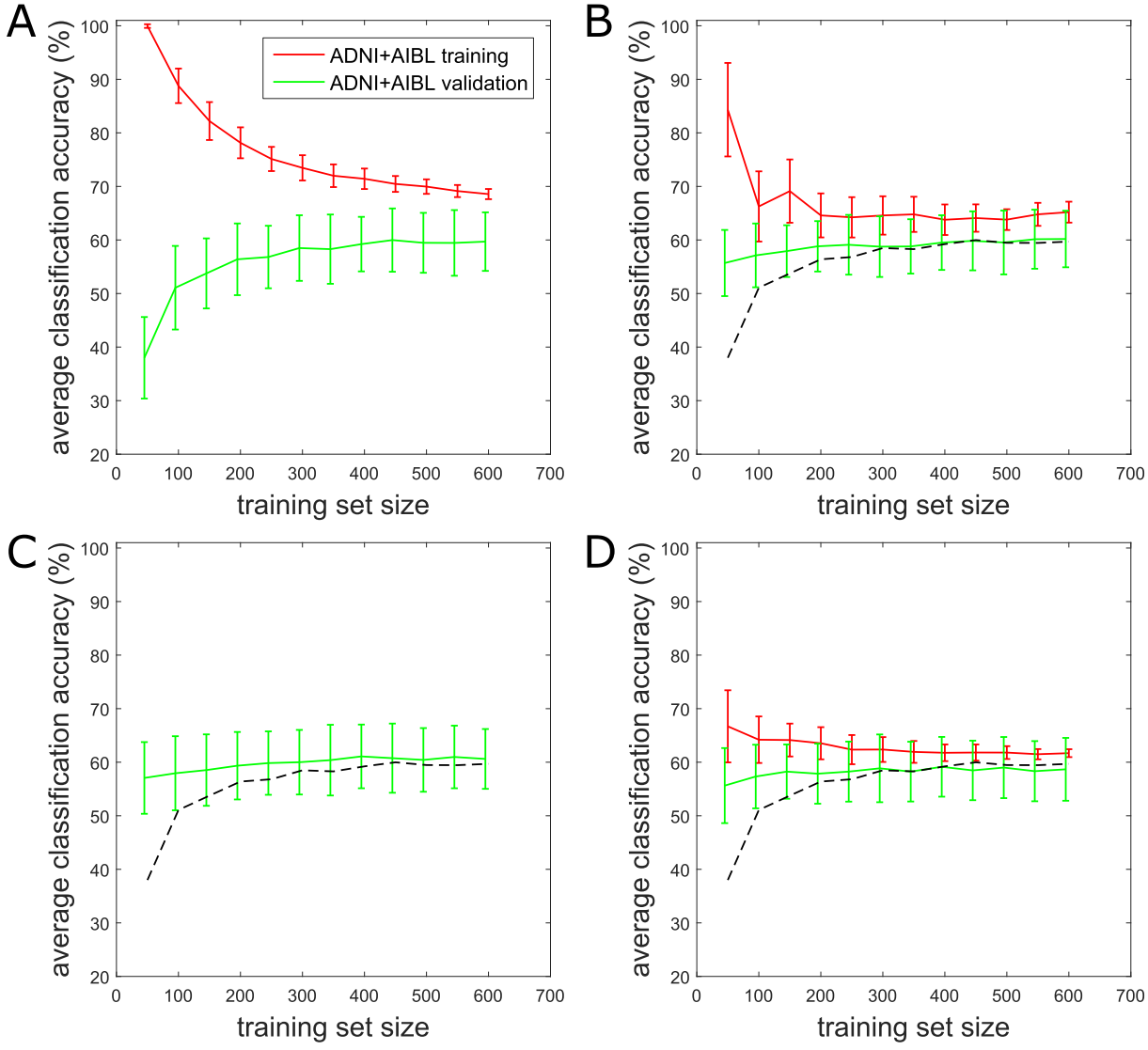

Fig. 3. Learning curves with MRI biomarkers as features. (A) LDA (the classifier used in the CADDementia challenge). (B) SVM with a radial Gaussian kernel. (C) Random forest classifier. Note that the training curve is not shown here because the training CA by design is ≈100%. (D) kNN classifier. Error bars mark ±standard deviation. The dashed black curve corresponds to the mean ADNI + AIBL validation accuracy of the LDA classifier, i.e., the green curve in (A).

Apart from our method, two other top-performing methods in the CADDementia challenge also used a combination of diverse MRI biomarkers (Bron et al., 2015), highlighting the benefit of combining different information. Wachinger et al. (2014) used a combination of volume, cortical thickness, and shape of different brain ROIs, and Ledig et al. (2014) used a combination of volume and cortical features including thickness in specific ROIs and intensity and texture in data-driven areas of the brain. The benefit of combining diverse MRI biomarkers was further emphasized by the five methods proposed by Ledig et al. (2014). Four submissions were based on the aforementioned four different types of features, and a fifth submission combined all four types of features. The combination of all four types of features proved to perform best, both in the challenge (Bron et al., 2015) and on ADNI data (Ledig et al., 2014).

In order to improve the results obtained using our method, more features would be needed. This could potentially be extraction of more information from the MRI scan. For example, extending texture and shape computation to other regions of the brain than the hippocampus. Texture and shape of other regions have already been successfully applied in AD (Chincarini et al., 2011; Tang et al., 2014), and in the CADDementia challenge, Wachinger et al. (2014) achieved good performance using the shape of as many as 44 different ROIs and Ledig et al. (2014) extracted texture from the entire brain. However, it may also be the case that the performance achieved among the top performing methods in the CADDementia challenge are at the limit of what a single structural T1-weighted MRI scan is capable of for this type of problem. Another option would be to include imaging biomarkers based on other MRI modalities such as fluid-attenuated inversion recovery or gradient-echo sequences to quantify vascular lesions that may coexist alongside AD pathology (Zekry et al., 2002).

Many methods in the CADDementia challenge, including our method, achieved a low true positive fraction for MCI (Bron et al., 2015). MCI is, in general, difficult to classify because it is a heterogeneous group in between NC and AD. The difference between true positive fractions for our method for the CADDementia test dataset (28.7%) compared with ADNI + AIBL (57.8%) further indicates that the MCI group in the challenge was particularly difficult to discriminate. Moreover, the combination of the low true positive fraction and the tendency for subjects with MCI to be misclassified as NC (Bron et al., 2015) indicate that CADDementia contained early MCIs, which potentially increased the difficulty of the classification problem because structural MRI is a late AD biomarker (Jack et al., 2013). Core biomarkers of AD that can detect the disease earlier than structural MRI directly measure abnormal protein build-up in cerebrospinal fluid (CSF) or use positron-emission tomography (PET) (Jack et al., 2013). There is, therefore, great potential in combining our MRI biomarker with these biomarkers for improved discrimination, especially of NC and MCI. Unfortunately, these biomarkers are more invasive, are not easily accessible, and are in some cases more costly, when compared with MRI, and blood-based biomarkers of protein build-up (Henriksen et al., 2014) may be a more viable candidate for combination with our MRI biomarker in the future.

The true positive fraction of MCI could be increased in our method by increasing the prior for MCI in the LDA. This would push the decision boundary between NC and MCI in the direction of NC with to effect of more correctly classified MCIs. This may, however, deteriorate the overall performance as exemplified by the second score we produced for the CADDementia challenge where the LDA priors were adapted to balance the confusion matrix in the CADDementia validation dataset (Sørensen et al., 2014).

Two versions of the hippocampal volume, FreeSurfer's estimate and NL patch, were included as features in the method for increased robustness for this central MRI imaging biomarker in AD. The hippocampus is affected early and severely in the AD pathological process (Braak and Braak, 1991; West et al., 1994), and the volume of this brain structure is the most widely applied (Jack et al., 2011b) and only qualified (Hill et al., 2014) MRI imaging biomarker in AD. It turned out that only one of the two hippocampal volume estimates contributed, and which contributed depended on the classification scenario. NL patch was selected when the two-class scenario MCI vs. AD was considered in the feature selection experiment while the FreeSurfer estimate was selected in the other scenarios. This should, however, be interpreted in conjunction with the other selected features, since restricting the algorithm to use only one of the hippocampal volume estimates as feature produced the following AUCs (reported as NC vs. MCI vs. AD/NC vs. MCI/NC vs. AD/MCI vs. AD): FreeSurfer 0.74/0.77/0.89/0.65 and NL patch 0.70/0.74/0.85/0.61.

4. Conclusions

We propose the combination of volumetry, cortical thickness, hippocampal shape, and hippocampal texture for differential diagnosis of NC, MCI, and AD using a single T1-weighted structural MRI scan. The combination of such diverse MRI biomarkers resulted in a multi-class CA of 62.7% on publicly available reference datasets (ADNI and AIBL). A similar CA of 63.0% was achieved in the CADDementia challenge, which resulted in a first place in the competition. The forward feature selection experiments revealed that hippocampal texture was the most important feature in the algorithm followed by hippocampal volume, ventricular volume, and parietal lobe thickness. The learning curve results, along with the fact that other challenge participants using relatively similar features and training data did not surpass our performance, indicate that other types of information are needed in order to improve beyond the obtained performance for this type of problem, for example, other MRI modalities or non-MRI core biomarkers of AD.

In summary, this paper

- describes and analyzes the winning algorithm in the CADDementia challenge;
- shows the importance of hippocampal texture as a feature in the algorithm; and
- conjectures that additional (possibly non-structural MRI) features are needed in order to significantly improve diagnostic performance.

Disclosures

M. Lillholm and M. Nielsen are shareholders in Biomediq A/S. The remaining authors report no disclosures.

Acknowledgments

This work was supported in part by the Danish National Advanced Technology Foundation (project 034-2011-5, "Early MRI diagnosis of Alzheimer's Disease") and in part by Eurostars (project 8234, "MR Brain Image Quantification in Dementia").

ADNI acknowledgments: ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated

by the Foundation for the National Institutes of Health. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

CADDementia acknowledgments: Data used in the preparation of this article were obtained from the CADDementia challenge (http://caddementia.grand-challenge.org).

References

Achterberg, H.C., van der Lijn, F., den Heijer, T., Vernooij, M.W., Ikram, M.A., Niessen, W.J., de Bruijne, M., 2014. Hippocampal shape is predictive for the development of dementia in a normal, elderly population. Hum. Brain Mapp. 35, 2359–2371. http://dx.doi.org/10.1002/hbm.22333.

Alkoot, F.M., Kittler, J., 2002. Moderating k-NN classifiers. Pattern Anal Appl 5, 326–332.

Anker, C., 2014. Segmentation of Subcortical structures in T1 weighted MRI as a component of a Brain Atrophy Computation Pipeline. Master's thesis. Technical University of Denmark, Department of Applied Mathematics and Computer Science.

Anker, C., Pai, A., Sørensen, L., Lyksborg, M., Conradsen, K., Larsen, R., Nielsen, M., 2014. Automated hippocampal segmentation using new standardized manual segmentations from the Harmonized Hippocampal Protocol. Alzheimers Dement 10, 415–P416. Supplement.

Ashburner, J., Friston, K.J., 2000, Jun. Voxel-based morphometry-the methods. Neuroimage 11, 805–821. http://dx.doi.org/10.1006/nimg.2000.0582.

Binnewijzend, M.A.A., Kuijer, J.P.A., Benedictus, M.R., van der Flier, W.M., Wink, A.M., Wattjes, M.P., van Berckel, B.N.M., Scheltens, P., Barkhof, F., 2013, Apr. Cerebral blood flow measured with 3D pseudocontinuous arterial spin-labeling MR imaging in Alzheimer disease and mild cognitive impairment: a marker for disease severity. Radiology 267, 221–230. http://dx.doi.org/10.1148/radiol.12120928.

Boccardi, M., Bocchetta, M., Morency, F.C., Collins, D.L., Nishikawa, M., Ganzola, R., Grothe, M.J., Wolf, D., Redolfi, A., Pievani, M., Antelmi, L., Fellgiebel, A., Matsuda, H., Teipel, S., Duchesne, S., Jack, C.R., Jr, Frisoni, G.B., EADC-ADNI Working Group on The Harmonized Protocol for Manual Hippocampal Segmentation and for the Alzheimer's Disease Neuroimaging Initiative, 2015. Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. Alzheimers Dement 11, 175–183.

Braak, H., Braak, E., 1991. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol 82, 239–259.

Bron, E.E., Smits, M., van der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Orellana, C.M., Meijboom, R., Pinto, M., Meireles, J.R., Garrett, C., Bastos-Leite, A.J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cárdenas-Pena, D., Álvarez Mezao, A.M., Dolph, C.V., Iftekharuddin, K.M., Eskildsen, S.F., Coupé, P., Fonov, V.S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. Neuroimage 111, 562–579.

Chincarini, A., Bosco, P., Calvini, P., Gemme, G., Esposito, M., Olivieri, C., Rei, L., Squarcia, S., Rodriguez, G., Bellotti, R., Cerello, P., De Mitri, I., Retico, A., Nobili, F., for the Alzheimer's Disease Neuroimaging Initiative, 2011, Sep. Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease. Neuroimage 58, 469–480.

Colliot, O., Chételat, G., Chupin, M., Desgranges, B., Magnin, B., Benali, H., Dubois, B., Garnero, L., Eustache, F., Lehéricy, S., 2008. Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus. Radiology 248, 194–201. http://dx.doi.org/10.1148/radiol.2481070876.

Cortes, C., Vapnik, V., 1995, Sep. Support-vector networks. Mach Learn 20, 273–297.

Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Collins, D.L., A.d.N.I., 2012. Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease. Neuroimage 59, 3736–3747.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. Neuroimage 54, 940–954. http://dx.doi.org/10.1016/j.neuroimage.2010.09.018.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., for the Alzheimer's Disease Neuroimaging Initiative, 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 56, 766–781.

Davatzikos, C., Resnick, S.M., Wu, X., Parmpi, P., Clark, C.M., 2008. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. Neuroimage 41, 1220–1227. http://dx.doi.org/10.1016/j.neuroimage.2008.03.050.

Ellis, K.A., Bush, A.I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N.T., Lenzo, N., Martins, R.N., Maruff, P., Masters, C., Milner, A., Pike, K., Rowe, C., Savage, G., Szoeke, C., Taddei, K., Villemagne, V., Woodward, M., Ames, D., the AIBL Research Group, 2009. The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int Psychogeriatr 21, 672–687.

Ellis, K.A., Rowe, C.C., Villemagne, V.L., Martins, R.N., Masters, C.L., Salvado, O., Szoeke, C., Ames, D., the AIBL Research Group, 2010. Addressing population aging and Alzheimer's disease through the Australian imaging biomarkers and lifestyle study: collaboration with the Alzheimer's Disease Neuroimaging Initiative. Alzheimers Dement 6, 291–296.

Eskildsen, S.F., Coup, P., Garca-Lorenzo, D., Fonov, V., Pruessner, J.C., Collins, D.L., for the Alzheimer's Disease Neuroimaging Initiative, 2013. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. Neuroimage 65, 511–521. http://dx.doi.org/10.1016/j.neuroimage.2012.09.058.

Falahati, F., Westman, E., Simmons, A., 2014. Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. J Alzheimers Dis 41, 685–708. http://dx.doi.org/10.3233/JAD-131928.

Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc Natl Acad Sci USA 97, 11050–11055.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Gerardin, E., Chtelat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.-S., Niethammer, M., Dubois, B., Lehricy, S., Garnero, L., Eustache, F., Colliot, O., for the Alzheimer's Disease Neuroimaging Initiative, 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage 47, 1476–1486.

Gower, J., 1975. Generalized procrustes analysis. Psychometrika 40, 33–51.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second, Springer-Verlag.

Henriksen, K., OBryant, S.E., Hampel, H., Trojanowski, J.Q., Montine, T.J., Jeromin, A., Blennow, K., Lönneborg, A., Wyss-Coray, T., Soares, H., Bazenet, C., Sjögren, M., Hu, W., Lovestone, S., Karsdal, M.A., Weiner, M.W., for the Blood-Based Biomarker Interest Groupy, 2014. The future of blood-based biomarkers for Alzheimer's disease. Alzheimers Dement 10, 115–131.

Hill, D.L.G., Schwarz, A.J., Isaac, M., Pani, L., Vamvakas, S., Hemmings, R., Carrillo, M.C., Yu, P., Sun, J., Beckett, L., Boccardi, M., Brewer, J., Brumfield, M., Cantillon, M., Cole, P.E., Fox, N., Frisoni, G.B., Jack, C., Kelleher, T., Luo, F., Novak, G., Maguire, P., Meibach, R., Patterson, P., Bain, L., Sampaio, C., Raunig, D., Soares, H., Suhy, J., Wang, H., Wolz, R., Stephenson, D., 2014. Coalition against major diseases/European medicines agency biomarker qualification of hippocampal volume for enrichment of clinical trials in predementia stages of Alzheimer's disease. Alzheimers Dement 10, http://dx.doi.org/10.1016/j.jalz.2013.07.003. (421–9.e3).

Igel, C., Heidrich-Meisner, V., Glasmachers, T., 2008. Shark. J Mach Learn Res 9, 993–996.

Jaakkola, T., Diekhans, M., Haussler, D., 1999. Using the Fisher kernel method to detect remote protein homologies. ISMB. AAAI Press., pp. 149–158.

Jack, C.R., Jr, Albert, M.S., Knopman, D.S., McKhann, G.M., Sperling, R.A., Carrillo, M.C., Thies, B., Phelps, C.H., 2011a. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 7, 257–262. http://dx.doi.org/10.1016/j.jalz.2011.03.004.

Jack, C.R., Jr, Barkhof, F., Bernstein, M.A., Cantillon, M., Cole, P.E., Decarli, C., Dubois, B., Duchesne, S., Fox, N.C., Frisoni, G.B., Hampel, H., Hill, D.L.G., Johnson, K., Mangin, J.-F., Scheltens, P., Schwarz, A.J., Sperling, R., Suhy, J., Thompson, P.M., Weiner, M., Foster, N.L., 2011b. Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. Alzheimers Dement 7, 474–485.e4. http://dx.doi.org/10.1016/j.jalz.2011.04.007.

Jack, C.R., Jr, Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S., Shaw, L.M., Vemuri, P., Wiste, H.J., Weigand, S.D., Lesnick, T.G., Pankratz, V.S., Donohue, M.C., Trojanowski, J.Q., 2013. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol 12, 207–216. http://dx.doi.org/10.1016/S1474-4422(12)70291-0.

Jain, A., Duin, R., Mao, J., 2000. Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 22, 4–37. http://dx.doi.org/10.1109/34.824819.

Klein, S., Loog, M., van der Lijn, F., den Heijer, T., Hammers, A., de Bruijne, M., van der Lugt, A., Duin, R., Breteler, M., Niessen, W., 2010. Early diagnosis of dementia based on intersubject whole-brain dissimilarities. Proceedings of IEEE International Symposium on Biomedical Imaging: from Nano to Macro. 2010. pp. 249–252.

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Jr, Ashburner, J., Frackowiak, R.S.J., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131, 681–689. http://dx.doi.org/10.1093/brain/awm319.

Knopman, D.S., DeKosky, S.T., Cummings, J.L., Chui, H., Corey-Bloom, J., Relkin, N., Small, G.W., Miller, B., Stevens, J.C., 2001, May. Practice parameter: diagnosis of dementia (an evidence-based review). report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology 56, 1143–1153.

Ledig, C., Guerrero, R., Tong, T., Gray, K., Schmidt-Richberg, A., Makropoulos, A., Heckemann, R.A., Rueckert, D., 2014. Alzheimers disease state classification using structural volumetry, cortical thickness and intensity features. MICCAI 2014 - Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data. pp. 5564.

Lillemark, L., Sørensen, L., Pai, A., Dam, E.B., Nielsen, M., for the Alzheimer's Disease Neuroimaging Initiative, 2014. Brain region's relative proximity as marker for Alzheimer's disease based on structural MRI. BMC Med Imaging 14:21,

Lindeberg, T., 2008. Scale-space. Wiley Encyclopedia of Computer Science and Engineering. John Wiley & Sons, Inc., pp. 2495–2504.
Manjón, J.V., Coupé, P., 2016. volBrain: an online MRI brain volumetry system. Front Neuroinform 10, 30. http://dx.doi.org/10.3389/fninf.2016.00030.
McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E.M., 1984. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA work group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 34, 939–944.
Næss-Schmidt, E., Tietze, A., Blicher, J.U., Petersen, M., Mikkelsen, I.K., Coupé, P., Manjón, J.V., Eskildsen, S.F., 2016. Automatic thalamus and hippocampus segmentation from MP2RAGE: comparison of publicly available methods and implications for DTI quantification. Int J Comput Assist Radiol Surg 11, 1979–1991. http://dx.doi.org/10.1007/s11548-016-1433-0.
Nyúl, L.G., Udupa, J.K., 1999. On standardizing the MR image intensity scale. Magn Reson Med 42, 1072–1081.
Ojala, T., Pietikäinen, M., Harwood, D., 1996. A comparative study of texture measures with classification based on featured distributions. Pattern Recognit 29, 51–59.
Perlich, C., Provost, F., Simonoff, J.S., 2003. Tree induction vs. logistic regression: a learning-curve analysis. J Mach Learn Res 4, 211–255.
Petersen, R.C., 2004. Mild cognitive impairment as a diagnostic entity. J Intern Med 256, 183–194. http://dx.doi.org/10.1111/j.1365-2796.2004.01388.x.
Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., Jack, C., Jr, Jagust, W.J., Shaw, L.M., Toga, A.W., Trojanowski, J.Q., Weiner, M.W., 2010. Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. Neurology 74, 201–209. http://dx.doi.org/10.1212/WNL.0b013e3181cb3e25.
Poulin, S.P., Dautoff, R., Morris, J.C., Barrett, L.F., Dickerson, B.C., for the Alzheimer's Disease Neuroimaging Initiative, 2011. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. Psychiatry Res 194, 7–13.
Ramani, A., Jensen, J.H., Helpern, J.A., 2006. Quantitative MR imaging in Alzheimer disease. Radiology 241, 26–44. http://dx.doi.org/10.1148/radiol.2411050628.
Sabuncu, M.R., Konukoglu, E., for the Alzheimer's Disease Neuroimaging Initiative, 2015. Clinical prediction from structural brain MRI scans: a large-scale empirical study. Neuroinformatics 13, 31–46.
Singh, V., Chertkow, H., Lerch, J.P., Evans, A.C., Dorr, A.E., Kabani, N.J., 2006. Spatial patterns of cortical thinning in mild cognitive impairment and Alzheimer's disease. Brain 129, 2885–2893. http://dx.doi.org/10.1093/brain/awl256.
Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 17, 87–97. http://dx.doi.org/10.1109/42.668698.
Sørensen, L., Igel, C., Liv Hansen, N., Osler, M., Lauritzen, M., Rostrup, E., Nielsen, M., for the Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing, 2016. Early detection of Alzheimer's disease using MRI hippocampal texture. Hum Brain Mapp 37, 1148–1161.
Sørensen, L., Nielsen, M., Lo, P., Ashraf, H., Pedersen, J., de Bruijne, M., 2012. Texture-based analysis of COPD: a data-driven approach. IEEE Trans Med Imaging 31, 70–78.
Mads, Sørensen, L., Pai, A., Anker, C., Balas, I., Igel, C., 2014. Dementia diagnosis using MRI cortical thickness, shape, texture, and volumetry. MICCAI 2014 - Challenge on Computer-aided Diagnosis of Dementia Based on Structural MRI Data. pp. 111–118.
Tanabe, J.L., Amend, D., Schuff, N., DiSclafani, V., Ezekiel, F., Norman, D., Fein, G., Weiner, M.W., 1997. Tissue segmentation of the brain in Alzheimer disease. AJNR Am J Neuroradiol 18, 115–123.
Tang, X., Holland, D., Dale, A.M., Younes, L., Miller, M.I., A.D.N.I., 2014. Shape abnormalities of subcortical and ventricular structures in mild cognitive impairment and alzheimer's disease: detecting, quantifying, and predicting. Hum Brain Mapp 35, 3701–3725.
Vapnik, V., 1998. Statistical Learning Theory. Springer, New York, USA.
Wachinger, C., Batmanghelich, K., Golland, P., Reuter, M., 2014. BrainPrint in the computer-aided diagnosis of Alzheimers disease. MICCAI 2014 - Challenge on Computer-aided Diagnosis of Dementia Based on Structural MRI Data. pp. 129–138.
Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., Morris, J.C., Petersen, R.C., Saykin, A.J., Schmidt, M.E., Shaw, L., Siuciak, J.A., Soares, H., Toga, A.W., Trojanowski, J.Q., for the Alzheimer's Disease Neuroimaging Initiative, 2012. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimers Dement 8, S1–S68. Supplement.
West, M.J., Coleman, P.D., Flood, D.G., Troncoso, J.C., 1994. Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. Lancet 344, 769–772.
Westman, E., Aguilar, C., Muehlboeck, J.-S., Simmons, A., 2013. Regional magnetic resonance imaging measures for multivariate analysis in Alzheimer's disease and mild cognitive impairment. Brain Topogr 26, 9–23. http://dx.doi.org/10.1007/s10548-012-0246-x.
Wyman, B.T., Harvey, D.J., Crawford, K., Bernstein, M.A., Carmichael, O., Cole, P.E., Crane, P.K., Decarli, C., Fox, N.C., Gunter, J.L., Hill, D., Killiany, R.J., Pachai, C., Schwarz, A.J., Schuff, N., Senjem, M.L., Suhy, J., Thompson, P.M., Weiner, M., Jack, C.R., Jr, for the Alzheimer's Disease Neuroimaging Initiative, 2013. Standardization of analysis sets for reporting results from ADNI MRI data. Alzheimers Dement 9, 332–337.
Zekry, D., Hauw, J.-J., Gold, G., 2002. Mixed dementia: epidemiology, diagnosis, and treatment. J Am Geriatr Soc 50, 1431–1438.